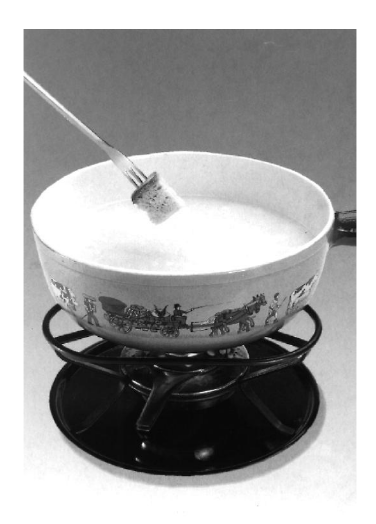
Des goûts et des couleurs ou Quel crédit accorder aux «tests» dans le secteur de la consommation ?[1]



Ce texte aurait tout aussi bien pu s'intituler «La généralisabilité et la fondue». En effet, l'idée des remarques et des analyses qui suivent nous est venue à la lecture d'une enquête menée il y a quelques mois par un quotidien vaudois. Cinq «spécialistes» (deux fromagers, une consommatrice, un journaliste, un ancien restaurateur) ont évalué cinq fondues. Deux d'entre elles avaient été tirées d'une boîte produite par l'industrie fromagère et apprêtées selon les règles de l'art; les trois autres avaient été servies par trois restaurants de la région. Chaque fondue a été évaluée selon cinq critères, notés sur 10: aspect, onctuosité, légèreté, persistance du goût et goût (cette dernière note étant comptée double). On a obtenu ainsi un score moyen censé permettre de classer les cinq produits testés et de comparer les productions industrielles *vs* artisanales.[2]

Mais, au fait, quelle confiance avoir dans ces évaluations, ce classement et ces comparaisons ? La question se pose chaque fois que nous lisons de tels «tests» dans les journaux, dans les magazines spécialisés ou que nous regardons à la télévision des émissions destinées aux consommateurs. Les enjeux de telles évaluations ne sont pas minces. Pour les commerçants d'abord, avec des conséquences financières non négligeables; pour les consommateurs ensuite, potentiellement influencés dans leurs habitudes d'achats.

Lorsqu'il s'agit de comparer la longévité de piles électriques ou la résistance aux chocs de casques pour motocyclistes, les associations de consommateurs ou les journaux font appel à des laboratoires spécialisés. Ceux-ci travaillent en se référant à des normes reconnues (de type ISO nnnn); l'information est fournie en principe dans l'article ou dans l'émission, et le test peut être refait selon le même protocole. Souvent rien de tel dans les «tests» auxquels nous faisions allusion plus haut. Tout semble se passer comme si on admettait que les aspects évalués (les goûts et les couleurs !) relevant de la subjectivité, il n'était pas nécessaire de recourir à un dispositif d'évaluation systématique et rigoureux; comme si le destinataire-

---

[1] Version revue et augmentée en 2004 d'un article paru dans le *Bulletin de l'ADMÉÉ* en 1998. (*Bulletin de l'ADMÉÉ* 97/3 – 98/1, pages 15 à 17).

[2] Je remercie mon collègue R. Capel (Université de Lausanne) de m'avoir transmis les informations sur lesquelles porte l'analyse qui suit.

consommateur n'avait pas besoin de renseignement sur ce dispositif pour estimer la fiabilité des résultats.

Pour revenir à notre exemple, un certain nombre de précautions devraient être prises dans un test de ce type. Les premières concernent naturellement le choix (l'échantillonnage) des fondues et des dégustateurs, les conditions de préparation ou de dégustation des fondues et les modalités des évaluations permettant d'éviter toutes sortes de biais bien connus. Par exemple, comment éviter une contamination entre les appréciations si les cinq dégustateurs se retrouvent autour du même caquelon ?

Un autre type de précaution, qui nous intéresse plus particulièrement ici, consisterait à tester la fiabilité du dispositif d'évaluation lui-même. A-t-il les qualités docimologiques nécessaires pour classer de façon fidèle les fondues, avec quelle marge d'erreur ? Permet-il par exemple de prétendre que la fondue préparée aux «Trois sifflets» (sic) avec sa moyenne de 8,3 est vraiment meilleure que la fondue industrielle de marque Migros (moyenne 5,4) ? Ce même dispositif pourrait-il donner une réponse fiable à la question fondamentale (au moins pour un Suisse) soulevée dans l'article: Vaut-il mieux aller manger la fondue au bistrot ou la préparer chez soi à partir d'une boîte ?[3] Serait-il d'autre part possible de fixer un seuil (par exemple la note 7 sur 10) au-dessus duquel on pourrait attribuer aux fondues un label d'excellence, avec une marge d'erreur raisonnable ?

Faute des précautions et des informations que nous venons d'esquisser, il nous apparaît *a priori* imprudent d'attribuer du crédit aux tests qui nous (pré)occupent. Mais peut-être sommes-nous exagérément pessimistes.

Nous avons eu l'occasion de le vérifier dans le cas gastronomique qui nous occupe. En effet, nous disposions du détail des évaluations, publié par le journal, soit un ensemble de 125 notes constitué par le croisement des 3 facettes Fondues (n = 5) x Évaluateurs (n = 5) x Critères (n = 5)[4]. Pour cette vérification, nous avons fait comme si nous étions dans une phase de test du dispositif et comme si les fondues et les évaluateurs avaient été choisis aléatoirement parmi un très grand nombre de fondues et de dégustateurs possibles. Pour le traitement statistique, nous avons eu recours au modèle de la généralisabilité et au logiciel *Etudgen*[5], qui s'impose dans

---

[3] Je préfère personnellement une troisième solution: préparer la fondue avec des fromages suisses soigneusement sélectionnés par moi. Cette modalité pourrait figurer dans un prochain test.

[4] Cf. le tableau des données en annexe 1. Dans cette analyse, nous avons renoncé à doubler la note attribuée au critère *goût*.

[5] Cf. à ce sujet les ouvrages: *Assurer la mesure*, de J. Cardinet & Y. Tourneur, P. Lang, Berne, 1985, et D. Bain & G. Pini: *Pour évaluer vos évaluations: la généralisabilité, mode d'emploi*, Centre de recherches psychopédagogiques du Cycle d'orientation, Genève, 1996. Au moment de la réédition du présent texte (2004), un nouveau logiciel, *EduG 2.0 français*, est disponible pour PC auprès de Dagmar Hexel, Service de la recherche en éducation, quai du Rhône 12, CH-1205 ; e-mail : Dagmar.HEXEL@etat.ge.ch.

une analyse complexe de ce type. Un premier plan de mesure (F/EC) considérait comme objets d'évaluation les Fondues (*facette de différenciation* aléatoire infinie) sans distinction d'origine (artisanale ou industrielle) et mettait sur la *face d'instrumentation* (= des moyens de mesure) les Évaluateurs (facette aléatoire infinie) et les Critères (facette fixée).

A notre grand surprise, le dispositif s'est révélé nettement plus fiable (généralisable) que nous ne le supposions: le *coefficient de généralisabilité relative* rhô carré (ρ^2 rel.) de 0.87 indique que l'on peut faire quelque crédit au classement établi par l'enquête (mesure relative). Le *coefficient de généralisabilité absolue*, lui aussi satisfaisant (ρ^2 abs. = 0.86), montre qu'on peut situer également avec une fiabilité satisfaisante les notes moyennes des fondues sur l'échelle d'évaluation de 1 à 10, par exemple par rapport à un seuil d'excellence comme la note 7.0 (mesure absolue)[6]. La valeur de l'indice ρ^2 relatif ou absolu passe en effet de 0 à 1 quand l'importance relative de la variance de différenciation (due ici aux différences entre fondues) augmente par rapport à la variance totale (variance de différenciation + variance d'erreur), et l'on considère le coefficient ρ^2 comme satisfaisant quand il est égal ou supérieur au seuil de 0.80.

L'avantage du modèle de la généralisabilité, seul applicable dans le cas de plans complexes comme ceux que nous traitons ici, réside aussi dans la possibilité d'aller plus loin dans l'analyse. Il permet d'expliquer par exemple la bonne fiabilité constatée par le fait que
- d'une part, le phénomène à évaluer (l'estimation de la qualité des fondues) est relativement contrasté: les notes moyennes vont de 5,4 à 8,3; d'où une variance de différenciation élevée;
- d'autre part, les appréciations des évaluateurs sont relativement convergentes: ils classent les différentes fondues *grosso modo* de la même façon et utilisent en moyenne la même zone de l'échelle pour leurs évaluations; d'où des variances d'erreurs relativement faibles (variance d'interaction FE, pour le coefficient relatif; variance d'interaction FE plus variance de E pour le coefficient absolu)[7].

L'analyse d'un autre test de ce type (vins blancs de Suisse romande testés par l'émission de TV «À Bon Entendeur» il y a quelques années) nous laisse supposer qu'on obtient des résultats relativement satisfaisants lorsqu'on fait appel à au moins 4 ou 5 dégustateurs spécialistes habitués à évaluer certains critères (d'où homogénéité relative des estimations). A condition naturellement de faire porter le test sur une

---

[6] Le lecteur intéressé par le détail des résultats trouvera en annexe 2 le listing des analyses pour le plan de mesure F/EC.

[7] Cf. annexe 2. La facette Critère étant fixée, ni elle ni ses interactions avec les autres facettes ne contribuent aux erreurs de mesure.

gamme assez variée de produits de qualités *a priori* très différentes. Le dispositif en l'état ne serait probablement pas fiable pour évaluer un ensemble relativement homogène réunissant uniquement des produits de haut de gamme (variance de différenciation faible par rapport aux erreurs de mesure).

Pour revenir à la comparaison des fondues, le calcul des marges d'erreurs (intervalles de confiance) permet d'attester la supériorité nette de la fondue la mieux classée, servie dans un restaurant, sur les deux dernières, préparées à partir d'une boîte. Le touriste de passage à Vevey aurait donc avantage, semble-t-il, à choisir le café des « Trois sifflets » pour tester la version vaudoise de cette spécialité suisse (publicité gratuite !). C'est d'ailleurs la seule fondue à laquelle on pourrait donner avec bonne conscience le label d'excellence mentionné plus haut (note significativement supérieure au seuil de 7 si l'on se réfère à l'*erreur absolue*).

En revanche, s'il s'agit de comparer non plus des fondues isolées mais les deux catégories: fondues industrielles *vs* artisanales, le plan risque de ne pas avoir la fiabilité suffisante. Nous avons testé ce cas de figure sur quatre des cinq fondues; le modèle d'analyse variance sur lequel s'appuie la généralisabilité exigeant un plan équilibré (2 fondues de chaque type), nous avons dû éliminer une des trois fondues artisanales par tirage au sort. Le plan d'observation comporte dans ce cas une quatrième facette (fixée) T: Type de fondue, dans laquelle sont incluses les Fondues elles-mêmes (F:T). Les deux coefficients de généralisabilité ainsi calculés (plan de mesure: T/FEC) n'atteignent pas tout à fait le seuil de 0.80 considéré traditionnellement comme satisfaisant (ρ^2 relatif et absolu = 0.724 et 0. 718).

Dans un tel cas, le modèle de la généralisabilité est précieux pour déterminer comment améliorer le dispositif d'évaluation – si possible à moindres frais – afin d'atteindre la fiabilité voulue. Une *analyse de facettes* (comparable à une analyse d'items classique) montre que la solution la plus économique (à tous points de vue!) serait d'éliminer l'évaluateur no 2: les ρ^2 relatif et absolu passeraient respectivement à 0.90 et 0.87. Entendez par là : il vaudrait mieux ne plus inviter à une telle dégustation un évaluateur de ce type (considéré en l'occurrence comme atypique). Encore faudrait-il pouvoir justifier, déontologiquement et docimologiquement, une telle décision. Le graphique de l'annexe 3 montre que ce dégustateur no 2 a un profil de notes moyennes différent des autres, du fait en particulier qu'il favorise dans son évaluation la fondue industrielle Gerber (Fondue ind. 2). Possédait-il des actions dans cette entreprise, avait-il forcé sur le blanc au moment de cette dégustation ou aurait-il été particulièrement conditionné au goût de cette fondue dès son jeune âge ? Ces hypothèses farfelues veulent mettre en évidence le fait que méthodologiquement il est nécessaire d'avoir de bonnes raisons pour éliminer tel ou tel niveau d'une facette considéré comme atypique.

Une analyse d'*optimisation* permet d'explorer d'autres pistes. Elle montre par exemple qu'il faudrait augmenter le nombre de fondues à au moins 3 de chaque type pour garantir une fiabilité satisfaisante (ρ^2 relatif et absolu = resp. 0.80 et 0.79) ; une telle solution pourrait être considérée comme pratiquement et économiquement raisonnable. En revanche, ce serait une mauvaise idée d'essayer d'augmenter le nombre d'évaluateurs : une cinquantaine de dégustateurs ne suffiraient pas à la tâche pour atteindre le seuil de 0.80.

Ces exemples (plus exemplatifs qu'exemplaires) d'exploration des possibilités d'amélioration illustrent tout l'intérêt heuristique et pratique du modèle de la généralisabilité quand il est utilisé dans une phase exploratoire visant à tester un dispositif d'évaluation.

Par ailleurs, si l'on considère le même problème de fiabilité du test non plus seulement d'un point de vue descriptif (rapport mesure / erreurs), mais également inférentiel, on constate que la marge d'erreur sur la composante de variance qui nous intéresse (F ou T) est considérable. Pour obtenir des valeurs plus satisfaisantes dans le cas du premier plan de mesure (F/EC), il faudrait au moins doubler le nombre de fondues testées, donc passer de 5 à 10 dégustations. Le modèle de la généralisabilité ne dit malheureusement pas si le foie et l'estomac des dégustateurs supporteraient ce surcroît de fromage.[8]

Daniel Bain
Centre de recherches psychopédagogiques du CO genevois[9]
novembre 1997

---

[8] Le lecteur qui voudrait refaire les analyses présentées brièvement ci-dessus trouvera les données nécessaires (sous forme de fichiers informatiques) sur la page *Exercices* du site Internet du groupe Edumétrie.

[9] Adresse de l'auteur en 2004 : route du Moulin-Roget 49, CH-1237 Avully.

Annexe 1 : Tableau des données

Fondue	**Type**	**Évaluateur 1**					**Évaluateur 2**					**Évaluateur 3**					**Évaluateur 4**					**Évaluateur 5**					**Moy.**
		asp.	onct.	lég.	per.	goût	asp.	onct.	lég.	per.	goût	asp.	onct.	lég.	per.	goût	asp.	onct.	lég.	per.	goût	asp.	onct.	lég.	per.	goût	
Migros	ind.	5	5	4	6	5	4	4	6	7	3	4	6	5	5	4	6	7	6	5	5	6	7	6	9	4	**5.36**
Marché, Aigle	art.	8	7	7	5	8	9	8	7	5	7	8	7	7	7	8	8	8	7	7	8	8	8	9	5	8	**7.36**
Bavaria, Montreux	art.	8	6	5	6	7	6	7	6	7	5	7	8	8	8	6	8	9	8	9	8	8	8	8	8	8.5	**7.30**
Gerber	ind.	8	8	8	6	4	8	9	8	10	8	5	4	6	6	5	8	8	6	6	5	6	7	5	7	3	**6.56**
3 sifflets, Vevey	art	7	9	8	8	7	8	9	9	10	9	6	6	7	8	8	9	7	8	9	9	10	8	10	10	9	**8.32**

Évaluateurs: 1 à 5

Types de fondues: ind. = industrielle art. = artisanale

Critères d'évaluation: asp. = aspect onct. = onctuosité lég. = légèreté per. = persistance du goût goût

Annexe 2 : Résultats des analyses de variance et généralisabilité pour le plan de mesure F/EC (listing produit par EduGf 2.0)

Test des fondues, article publié dans le Bulletin de l'ADMEE 1997/3

Plan d'observation et d'estimation

Facettes	Niveaux	Univers	Nom	Réduction
F	5	INF	fondues	
E	5	INF	évaluateurs - goûteurs	
C	5	5	critères d'évaluation	

Plan de mesure: F/EC

Analyse de Variance

Sources de var.	S.C.	D.L.	C.M.	Comp. aléat.	Comp. mixtes	Espér. mixtes	%	Erreurs types
F	121.0800	4	30.2700	0.9830	1.0567	1.0567	36.2	0.7019
E	22.4800	4	5.6200	0.0785	0.0707	0.0707	2.4	0.1401
C	9.2800	4	2.3200	-0.0095	-0.0095	-0.0076	0.0	0.0659
FE	61.6400	16	3.8525	0.5885	0.7705	0.7705	26.4	0.2588
FC	44.0400	16	2.7525	0.3685	0.3685	0.2948	10.1	0.1862
EC	11.4400	16	0.7150	-0.0390	-0.0390	-0.0312	0.0	0.0572
FEC	58.2400	64	0.9100	0.9100	0.9100	0.7280	24.9	0.1584

Coefficients de Généralisabilité

Plan de mesure: F/EC

Sources de variance	Variance de dif.	Sources de var.	Var. d'err. relative	Var. d'err. absolue	%
F	1.0567				
		E		0.0141	8.4
		C		0.0000	0
		FE	0.1541	0.1541	91.6
		FC	0.0000	0.0000	0
		EC		0.0000	0
		FEC	0.0000	0.0000	0
Totaux	1.0567		0.1541	0.1682	
Ecarts types	1.0280		0.3926	0.4102	

Coefficient de généralisabilité	relatif	absolu
	0.8727	0.8627

Moyenne générale pour les niveaux traités: 6.9800
Variance d'échantillonnage de la moyenne générale pour les niveaux traités: 0.2563

Annexe 3 : Profil des notes des évaluateurs pour 4 fondues
(ind. : industrielles ; art. : artisanales) ; plan d'observation à 4 facettes : T, F:T, E, C)

